DOCUMENT RESUME

ED 293 385 HE 020 546

AUTHOR Fong, Bobby
TITLE The External Examiner Approach to Assessment. AAHE Assessment Forum Paper.
INSTITUTION American Association for Higher Education, Washington, D.C.
SPONS AGENCY Association of American Colleges, Washington, D.C.
PUB DATE Jun 87
NOTE 27p.; Paper presented at the National Conference on Assessment in Higher Education (2nd, Denver, CO, June 14-17, 1987). Paper collected as part of the American Association for Higher Education Assessment Forum.
AVAILABLE FROM American Association for Higher Education, One Dupont Circle, Suite 600, Washington, DC 20036.
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Academic Standards; *Achievement Tests; Bachelors Degrees; College Students; Criterion Referenced Tests; *Educational Assessment; Higher Education; Knowledge Level; Majors (Students); Models; *Outcomes of Education; *Student Evaluation; Test Reliability; Test Validity
IDENTIFIERS *AAHE Assessment Forum; College Outcomes Assessment; *External Evaluation; Great Britain

ABSTRACT

The external examiner approach to assessment can address concerns that American educational standards are low, curricular coherence is lacking, and college students are learning insufficiently. The discussion: (1) contrasts the conditions in British and American higher education that make the British model inappropriate for the United States; (2) explores how the external examiner model nevertheless has applications that address American needs for assessment; (3) review the considerations of reliability and validity in external examiner use of comprehensive and oral examinations; and (4) gives examples of how American institutions are presently using external examiners to evaluate learning in courses, internships, and senior projects, as well as assessing summative learning in majors. (Author/KM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED293385

The External Examiner Approach To Assessment

by Bobby Fong

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
BOBBY
FONG
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

HE 020 546

THE AAHE ASSESSMENT FORUM

AMERICAN ASSOCIATION FOR HIGHER EDUCATION

THE EXTERNAL EXAMINER APPROACH TO ASSESSMENT

Bobby Fong

Assistant Project Director
for the
Association of American Colleges
program
"Assessing Learning in Academic Majors
Using External Faculty Examiners"

Paper commissioned by the
American Association for Higher Education
Assessment Forum
for the
Second National Conference on Assessment
in Higher Education
Denver
14-17 June 1987

Association of American Colleges
1818 R Street NW
Washington, D.C. 20009
(202) 387-3760

ABSTRACT

The external examiner approach to assessment can address concerns that American educational standards are low, curricular coherence is lacking, and students are learning insufficiently. The discussion 1) contrasts the conditions in British and American higher education that make the British model inappropriate for the United States; 2) explores how the external examiner model nevertheless has applications that address American needs for assessment; 3) reviews the considerations of reliability and validity in external examiner use of comprehensives and oral examinations; and 4) gives examples of how American institutions are presently using external examiners to evaluate learning in courses, internships, and senior projects, as well as assessing summative learning in majors.

One Dupont Circle
Suite 600
Washington, D.C. 20036
202/293-6440

AMERICAN ASSOCIATION
FOR HIGHER EDUCATION

Board of Directors

Chair
Joseph F. Kauffman
University of Wisconsin
Madison

Chair-Elect
Adele S. Simmons
Hampshire College

Vice Chair
Reatha Clark King
Metropolitan State University

Past Chair
Harriet W. Sheridan
Brown University

Carlos H. Arce
NuStats, Inc.

Estela M. Bensimon
Teachers College
Columbia University

Anne L. Bryant
American Association of
University Women

Donald L. Fruehling
McGraw-Hill, Inc.

Ellen V. Futter
Barnard College

Jerry G. Gaff
Hamline University

Zelda F. Gamson
University of Michigan

Stephen R. Graubard
Daedalus

Joseph Katz
State University of New York
at Stony Brook

Arthur E. Levine
Bradford College

Frank Newman
Education Commission
of the States

Alan Pifer
Carnegie Corporation
of New York

W. Ann Reynolds
The California State
University

Piedad F. Robertson
Miami-Dade Community
College

D. Wayne Silby
Groupware Systems

P. Michael Timpane
Teachers College
Columbia University

President
Russell Edgerton

The AAHE ASSESSMENT FORUM is a three-year project supported by the Fund for the Improvement of Postsecondary Education. It entails three distinct but overlapping activities:

--an annual conference
(the first scheduled for June 14-17, 1987, in Denver)

--commissioned papers
(focused on implementation and other timely assessment concerns; available through the Forum for a small fee)

--information services
(including consultation, referrals, a national directory, and more)

This paper is part of an on-going assessment collection maintained by the Forum. We are pleased to make it more widely available through the ERIC system.

For further information about ASSESSMENT FORUM activities, contact Patricia Hutchings, Director, AAHE ASSESSMENT FORUM, One Dupont Circle, Suite 600, Washington, DC 20036

THE EXTERNAL EXAMINER APPROACH TO ASSESSMENT

Calling upon an authority from outside the classroom to evaluate student learning is as old as American higher education. At colonial Harvard, according to Lowell (1912, p. 583):

> In 1650, . . . an order of the Overseers provided that there should be three weeks of visitation yearly, . . . that the yearly progress and sufficiency of scholars might be manifest. During these periods the students were to sit in the hall, to be examined by all comers in the Latin, Greek and Hebrew tongues, and in Rhetoric, Logic and Physics; "and they that expect to proceed bachelors that year, to be examined of their sufficiency according to the laws of the college"; any of them who were found insufficient in the judgment of any three of the visitors, being Overseers of the College, to be deferred to the following year.

In later years the Overseers appointed committees of visitors to examine in their stead, but this system of certifying learning persisted into the mid-nineteenth century and had parallels in many other American colleges.

The arrangement, however, proved increasingly inadequate as the assumption that there should be a core of learning common to all graduates broke down. The rise of scientific and technological studies, the advent of academic departments, and the institution of electives made the individual course and its examinations the most convenient way to assess student achievement (Rudolph, 1977). Thus the classroom teacher was accorded the responsibility for both teaching and certifying students.

By contrast, nineteenth-century Oxford and Cambridge made external examiners the keystone of a new system of written examinations. Tutors in the individual colleges were responsible for instruction, but the university certified prospective graduates in each department through a series of institution-wide essay examinations formulated and graded by tutors external to the college. The quality of teaching and learning was verified by a jury of peers. The success of this arrangement led to the establishment of similar systems in other universities and polytechnics and for secondary schools (Roach, 1971). External examiners remain central to the British educational process today.

Unlike the American approach, in which a major at times seems but a convenient aggregation of courses, the British system of comprehensive subject examinations ostensibly impels students to

strive for synthesis and tutors to convey some sense of coherence in their subject. The modern examiner is far more than a tester, for the office is responsible for maintaining standards and providing counsel for the improvement of instruction. Writes Omolewa (1978, p. 19):

> The external examiner rightly examines the institution including the teacher and the students. He usually offers suggestions on the development of the curriculum in a Department, and comments on the quality of students' scripts, dissertation or thesis. Existing external examiners' Reports refer to the standard of performance of students, their level of understanding of the subject and the suitability of the content of programmes of study.

Given that the concurrent American interest in assessment stems from concerns that educational standards are low, curricular coherence is lacking, and students are learning insufficiently, what can we gain from the external examiner approach? I want to approach this question in four ways: 1) by contrasting the conditions in British and American higher education that make the British model inappropriate for the United States; 2) by exploring how the external examiner model nevertheless has applications that address American needs for assessment; 3) by reviewing the problems of reliability and validity in external use of comprehensives and oral examinations; and 4) by giving examples of how American institutions are presently using external examiners to evaluate learning in courses, internships, and senior projects, as well as assessing summative learning in majors.

Perspectives across the water

According to the code of practice in the "External Examiner System for First Degree and Taught Master's Courses" (Committee of Vice Chancellors and Principals, 1984):

> The purposes of the external examiner system are to ensure, first and most important, that degrees awarded in similar subjects are comparable in standard in all universities in the United Kingdom, though their content does of course vary; and secondly that the assessment system is fair and is fairly operated in the classification of students.

The "first and most important" purpose of the British examiner system, to establish normative standards for the national university system, presumes a unitary conception of excellence. In theory, standards for awarding degrees should remain consistent from university to university, from subject to subject, and from year to

year. This has not been true in practice (Dale, 1959), but variations in honours degree results are considered undesirable departures from the ideal functioning of the system (Williams, 1979). Skeptics have suggested that what consistency there is may actually result from the efforts of examiners to norm each graduating class rather than from uniformity in student quality (Agar and Weltman, 1967; Oppenheim, et al., 1967). In any event, however, the "notion of the unity of standard implied by a British degree is deeply imbedded in the system" (Piper, 1985, p. 332).

It is difficult to foresee American higher education agreeing to a unitary conception of excellence, for a byword of the American system is "diversity." British higher education is still a "narrow gate," through which only a minority of the college-age population is permitted to enter. Those admitted are presumably the best-prepared as certified by batteries of entrance exams, also composed and marked by external examiners. The United States, by contrast, has been committed to equity in educational opportunity, a "wide gate" through which a majority of the college-age population enters to pursue some post-secondary learning. Alexander Astin (1985) has argued persuasively that American education operates under multiple definitions of excellence in order to serve a clientele that ranges from the academically accomplished to the marginal. Assisting students to fulfill whatever potential they have is a far different vision for higher education than the British ideal of unitary expectations for all its graduates. Any American assessment effort, including use of external examiners, must acknowledge the diverse missions and standards of our educational system.

The second purpose mentioned in the external code of practice is fairness in its operation. This is particularly essential given the weight carried by these examinations. The performance of students in the final year battery constitutes the main, and in some cases the sole, criterion of what they have learned over their matriculation. Although British universities are increasingly giving more credit for tutorial essays, projects, and other work, the examinations remain the primary evidence of summative learning. The pressures on the student are intense, for failure to obtain a proper degree of honors, or to pass at all, can mark one for life, since first or high second class honors are prerequisite for further study or prestigious employment. Concern over the reliability of examiner ratings is more than a matter of test design: the effects of an examiner mark would be comparable in America to a public announcement of one's IQ in the 1950's.

American higher education's use of "continuous assessment" (denoting the progressive evaluation via papers and exams by numbers of professors over a student's career) stands in stark contrast to the single hurdle of the British system. American students and faculty, I suspect, would be loathe to give up the grades and credits earned from course to course in favor of one summative experience at

the end of study. The British are bemused by continuous assessment. There is some concern that constantly studying for tests and writing graded papers would distract students from concentrated reading and thinking in a subject, discourage independent work, and substitute constant tension for the anxiety that a British student might feel only toward the end of a program (Cockburn and Ross, 1980). Americans might counter that continuous assessment spurs learning, allows for frequent feedback to students, and offers multiple opportunities for students to demonstrate achievement, rather than pointing toward a single juncture in one's career. Assessing summative learning may be a valuable addition to the knowledge we now have of American student achievement, but it will not replace course work and course evaluation.

A third purpose of the British examiner system grows out of the first two: if standards are to be kept constant and uniform, if students' futures depend on their performance on these examinations, then the syllabi of courses and tutors must accord with what will be examined. Therefore externals are encouraged to be general advisors to subjects which they will examine. They must maintain standards for programs as well as standards for student learning. In a recent survey, about half of the examiners had some involvement in course design at local institutions, that is, consulting on content and/or methods (Piper, 1985).

Such a regulatory function for examiners would be seen by American professors to impinge on their traditional prerogative to choose what and how to teach. Outside consultants are used in curriculum reviews and departmental self-studies, but only as they are invited at the behest of the faculty. Any assessment system that aims at improving instruction must be collegial in nature, respecting the autonomy exercised by American professors and their final control of their courses.

While the use of external examiners, modified for American conditions but informed by the British experience, can help address certain concerns in the current assessment debate, the American realities of institutional diversity, continuous assessment, and faculty autonomy militate against wholesale emulation of the British system. Indeed, these realities constitute limits to all schemes of assessment for American higher education, not just the external examiner approach. The particular danger of state or nationally-normed instruments imposed by legislatures or accrediting agencies is that they may ignore the uniqueness of institutions, overlook the evidence of cumulative student achievement, and relegate faculty involvement to the periphery. Such efforts may fail to accurately gauge learning or improve instruction. In principle, to be successful, any approach to assessment must involve faculty input and direction, respect existing indicators of what students are learning, and be mindful of the particular missions and clientele of the individual institution.

The external examiner approach

Given the constraints already discussed, what promise does the external examiner offer to American education? Seven potential advantages warrant discussion. The external examiner approach:

1) Separates instruction and evaluation

Joseph P. O'Neill (1983, p. 71) writes:

> . . . there is a conflict of interest in the way in which American colleges and universities certify instruction. . . . Faculty members not only teach but in effect guarantee, first, that their teaching meets established standards in both content and quality and, second, that students have learned what faculty have taught. There is no external mechanism to verify the integrity of the baccalaureate degree.

Such verification was for a long time not necessary because it was not demanded. Employers were content with the products of higher education, parents satisfied that their investment was well-spent, and students sanguine that they had been sufficiently equipped to meet the challenges that awaited them after graduation. The current calls for accountability, although led by legislative bodies and accrediting agencies, are in truth but the formalized discontents of the workplace, of parents, and of students. The demand is that learning ought to be demonstrable to people beyond those who teach the courses. This does not ignore the informed judgment of instructors, but it does ask that such judgment be corroborated by other means. External examiners offer one way of certifying learning without being open to charges of conflict of interest since they can assess student achievement without having a departmental or institutional stake in the outcome.

2) Examines for learning beyond the course

What is learned should have meaning beyond the duration of a course. Knowledge and skills have relevance to other courses in the curriculum, to the overall shape and thrust of the baccalaureate experience, to the particular occupations pursued by students after graduation, and to the general texture of social relations and citizenship in a community. This is not to disparage learning for its own sake, but to insist that true learning should overflow the artificial construct of the course. The idea of inviting an external examiner to help evaluate the learning of a student, whether in a given course or over the entire expanse of four years, is based on the assumption that what has been learned should make sense to an authoritative outsider conversant in the subject or discipline but not directly involved in the particular instruction of the student. This presumes that instruction was properly directed to knowledge and

skills commonly considered current and valuable by professional peers, and that the student can demonstrate mastery of such learning.

3) Necessitates faculty participation in the process

The use of examiners requires the continuing involvement of the faculty in selecting, orienting, and employing knowledgeable outsiders. The authority of the approach lies in the people chosen, rather than in the instruments they use. If examiners are to validate student learning, they first must be accorded legitimacy by the host faculty. The selection and use of externals in the American context necessarily depends on recommendations and referrals from the instructional staff. In turn, the host faculty's investment in the process should permit them to seriously entertain the examiners' comments and suggestions as to where the preparation of students as a whole has been strong or weak.

We should not underestimate the importance of collegial relations in establishing, maintaining, and verifying standards. Standards in disciplines are set in large part by the implicit culture of each profession. Writes O'Neill (1983, pp. 74-5):

> . . . an informal but powerful consensus. . .links individual departments at the college level with their corresponding discipline-based associations. These links in turn influence not only publishers and the materials that they produce but foundations and federal agencies in the projects that they fund. This interactive flow of information and attitude affects the way in which individual faculty members select and present course material, and it helps them to define the kinds of learning that they expect from students.
>
> Informal standards imposed by peer pressure presuppose that individual faculty members are in direct and relatively constant contact with others in their discipline. . . . An informal system of standards begins to break down when the crucial elements of personal contact and peer pressure fail to operate.

In this regard, use of external examiners necessitates systematic and periodic conversation with host faculty as to what constitutes current knowledge and whether the students exhibit these learnings. Collegial respect for each other's authority permits frank talk aimed at instructional improvement without such advice being dismissed as either uninformed or self-interested.

4) Considers distinctiveness of institutional programs

A test instrument, unless locally devised or locally modified, may not take into account the particular program emphases of an individual institution. A department may be responsible for imparting skills and knowledge common everywhere to any competent graduate of a discipline, but student clientele, teaching interests, regional commitments, and a host of other factors will lead to variations in program emphases and modes of instruction. No consultant or examiner can assess a program or students without some knowledge of the institutional context. A well-designed examiner system must include some orientation to what a given course or major comprises, including departmental goals, syllabi, requirements of the concentration, and student profiles. The object is not to bring the examiner into sympathy with the status quo, but to give the external some sense of how learning common to a subject anywhere is being locally carried out. It is here that assessing student outcomes is linked to assessing program instruction.

5) Permits feedback

The unfortunate aspect of multiple-choice examinations is that, without item analyses, which are frequently unavailable from commercial publishers, they provide little in the way of useful information as to where programs are strong and weak. Ranking students against a national norm does not in itself reveal where instruction has been deficient or lacking. By contrast, the apparently more cumbersome examination tools--written essays, oral interviews, and observations--tend to highlight patterns of strengths and weaknesses that can be readily perceived and enunciated by the external. A general impression of a program inevitably arises from an examiner's evaluation of a student cohort. If the goal of assessment is simply external accountability, then an appropriate score on a standardized examination may be sufficient, but for improvement of learning and instruction, it seems advantageous to entertain formats that can yield disinterested but knowledgeable analyses of how students perform.

6) Balances criterion and selection-referenced goals for instruction

John Harris (1986, p. 16) provides this distinction between selection and criterion-referenced tests:

> The focus of selection tests has not been to compare a student's performance to an absolute standard of knowledge or skill, but to the performance of others. The scoring and scaling methods of selection tests are intended to maximize individual differences for purposes of comparison.

> In contrast, the historic intent of educational tests is to determine how much of a body of knowledge one knows, or how skillful one is as compared to some pre-set standard. In more recent years, psychologists have referred to these as criterion-referenced tests. . . .
>
> The selection test approach works well when the purpose is to spread individuals over a continuum. But it is awkward, to say the least, when the purpose is to certify a level of competence. . . . Nevertheless the commercially available achievement tests you will come across have been built, for the most part, on the selection model. . . . If you use a usual standardized test to compare possible gains in knowledge or skills, or to compare different instructional approaches, first ask the test publisher if the scores can be interpreted in a criterion-referenced way. If not, be prepared for the differences to be insignificant and do <u>not</u> assume that the lack of significant differences is completely attributable to ineffective instruction.

Instructors have two goals for their classes. First, they want to invest a class with knowledge and skills that will enable each student to be competent in the subject of study. Second, they seek to maximize the abilities of individual students. Effectively, instructors seek a floor to student achievement, the pass-fail line, and above that to distinguish whether degrees of achievement are excellent, good, or merely satisfactory. Instruction in the classroom is guided by both criterion and selection-referenced considerations. Unlike commercial examinations, external examiners, themselves most frequently instructors at other institutions, are experienced in balancing these two goals in evaluation. An external's sense of what constitutes adequacy and excellence may differ from the host faculty's, but the discussion and negotiation of these standards is precisely where the perspective of the outsider becomes invaluable.

7) <u>Allows for testing of higher order cognitive skills</u>

The recent British literature on assessment has argued the utility of multiple-choice examinations for certain purposes. There is agreement that such "objective" tests are a superior means of sampling knowledge and comprehension, principally because the number of questions posed in the same time frame can be greater than is possible in essay formats. Nevertheless, there is a continuing belief that such an approach is less fitted to assess the higher order skills that Benjamin Bloom (1956) denominated in his taxonomy:

application, analysis, synthesis, and evaluation (Cockburn and Ross, 1981; Elton, 1982). Cox (1967) and Rowntree (1977) among others have suggested that progress from knowledge to evaluation can be likened to movement on a continuum from convergent to divergent problems. Convergent problems admit of one correct or appropriate answer. Divergent problems admit of multiple answers and problem-solving strategies. Thus, objective tests are inappropriate for evaluating higher order cognitive skills because their convergent format does not allow for divergent possibilities. On the other hand, the typical external examiner instruments, essays and orals, better permit students to demonstrate these skills.

Reliability and validity

Although external examiners may use a variety of instruments to assess student learning, the two modes traditionally associated with this approach are comprehensive examinations and oral interviews. Given the prominence of written comprehensives in England, the British literature has been especially concerned with the question of reliability, that is, the accuracy and stability of results. Since the classic study of this topic by Hartog and Rhodes (1935), research has confirmed that essay-type tests are beset by a number of variabilities. Writes Elton (1982, p. 110):

> The traditional type of examination in Britain, in which a student answers in 2-3 hours some few questions (typically 3-4 in arts subjects, rather more in the sciences) usually by choosing them from a larger number of questions offered, is worryingly unreliable. Not only do different examiners give different marks to the same candidate but they rank them in different order. . . .

Cox (1967) notes that examiners may differ in the range or dispersion of the marks they award. Others have observed that the same examiner grading the same set of tests on two different occasions may give different marks to the same paper (Eells, 1930; Elton, 1982). In addition to these variations in examiner reliability, there is the question of test-retest reliability, in which the same student may achieve widely different marks in successive administrations of a test. The explanations range from increased familiarity with the format to the effects of anxiety and what the candidate ate the night before. In light of such variations, Dale (1959) and Heywood (1977) cite evidence of how cumulative standards fluctuate from year to year and from subject to subject.

Such findings have led to a number of suggestions whereby reliability might be increased. One is to increase the number of markers so that averaging grades might reduce fluctuations. A second is to engage examiners in discussion of model answers and to break

questions down into elements that need to be addressed. This would alert examiners to possibilities and strategies not previously entertained. A third suggestion is to increase the number of questions to be answered while reducing or eliminating the choice of questions. Reducing choice increases the comparability of examinations. Increasing the number of questions allows greater sampling of knowledge and skills and results in a more uniform student performance over the exam. Unfortunately, increasing the number of questions would also affect the character of the essay examination since there would be less time for each answer. In a proposal to develop standardized open-ended questions, Warren (1977) has intimated that careful construction of questions, design of carefully-defined categories to which responses can be assigned, and use of multiple items (as few as six) may yield results with a high reliability about .90.

In addition are two recurrent proposals whereby comprehensive examination results would be supplemented by other measures. Dale (1959) and Cox (1967) discuss the desirability of using objective examinations, which when properly designed can have high reliability, to test for knowledge and comprehension. Such examinations could be an additional portion of the comprehensives presently in place and serve as a valuable check on the overall results. The second option, urged by Bassey (1971) and Elton (1982), among others, is to incorporate credit from coursework and projects into the final evaluation; in other words, to use continuous assessment. If adopted, both suggestions would move British and Commonwealth higher education closer to American practices.

At the same time, there is little likelihood that the British will abandon the use of comprehensive essay examinations, particularly in favor of objective tests. Considerations regarding test validity suggest that there are limitations to objective exams, however reliable they might be. Validity refers to whether a test measures what it is intended to measure, and its determination is dependent on context, the particular use made of the instrument. As Rippere (1974, p. 211) writes:

> . . .it is necessary to acknowledge one of the axioms of contemporary testing, namely, that a test may be put to a variety of uses, and that the sort of validity one seeks to demonstrate will depend on the particular way one is using the test. Results from the same test, used to draw different sorts of inferences, would need to be validated in ways appropriate to each type of inference. A test result validated for one purpose need not necessarily be valid for another. . . .

For example, the GRE area exam in literature in English may have predictive validity in distinguishing those English majors who will

do well in graduate school. However, given its intended clientele and its level of difficulty in terms of content and abilities examined as determined by the designers and experts, it may lack face validity as an exit examination to assess how well all senior English majors have mastered the discipline. Cronbach (1971, p. 447) notes, "One validates, not a test, but an interpretation of data arising from a specified procedure."

Mention was made earlier that the British believe objective examinations to be deficient in their capacity to measure higher order cognitive skills (cf. Cox, 1967; Elton, 1982). In Cronbach's taxonomy, this is a suggestion that objective examinations lack validity with regard to educational importance: "does the battery of measures neglect to observe any important outcome?" (1971, p. 446). The conviction of the British is that essay examinations, despite their relative unreliability compared with objective tests, nevertheless can have greater validity than objective tests as a measure of higher order cognitive skills. Writes Elton (1982 , pp. 115-6):

> Quite generally, the less predictable a learning outcome is the less reliably can it be measured, because the more it involves the examiners' judgment. Simple recall of knowledge, like the memorizing of a poem, can be checked objectively against the original knowledge, but anything that involves the higher mental abilities--application, analysis, synthesis and evaluation, in the terms of Bloom, can only be appraised in terms of the same abilities as practised subjectively by any individual examiner. It follows that the higher abilities, which are the ones which we normally wish to foster in higher education, by their very nature cannot be assessed as reliably as the lower ones. Hence there is an inescapable contradiction between high reliability and high validity in our assessment procedures. . . .
>
> The conclusion which we have reached, that reliability and validity can be traded off against each other, appears to contradict accepted wisdom in the psychometric literature which holds that it is not possible for a measurement to be valid unless it is reliable. This is correct as long as we are concerned with very high reliability, since any loss of reliability is found to lead to some loss of validity. In student assessment we are, however, inevitably dealing with much lower levels of reliability and the position here is quite different.

Rippere (1976, p. 213), argues along the same lines, noting:

> . . .the degree of validity required for some particular purpose depends on that purpose and the degree required for some other purpose might be higher or lower; and. . .for some purposes a fairly unreliable measure may suffice, and it might even be preferable to a more reliable measure if it costs less or is otherwise more convenient.

Elton concludes, ". . .in stressing the need for reliability, we run the risk of testing mainly that which can be tested reliably. Not only can this have a most deleterious effect on teaching and learning, but it can actually lead to valuable areas of learning being excluded from the curriculum" (p. 116). Elton and Heywood (1977) recognize a "backwash" effect to examinations, where tests influence modes of instruction and learning. This would be familiar to those of us who lament the deficient writing skills of seniors who have been required to take nothing but multiple choice examinations since their freshman composition course.

These concerns of the British were echoed by Norman Frederiksen of the Educational Testing Service (1984). In response to claims that it is possible to write multiple choice test items that probe for skills of application, analysis, or interpretation, Frederiksen was able to find only two studies of the matter, and both, involving commercially published exams, found that the majority of items required only recall of information. Furthermore, attempts in one case to rewrite the examination to test for complex cognitive processes still resulted in a majority of items that were judged to require only recall. Frederiksen reiterates British contentions that the multiple choice format itself militates against the demonstration of complex cognitive problem-solving skills that can be tested by free-response formats. He also voices his concern that as multiple choice tests drive out other testing procedures which might be used in school evaluation, "the abilities that are most easily and economically tested become the ones that are most taught. If educational tests fail to represent the spectrum of knowledge and skills that ought to be taught, they may introduce bias against teaching important skills that are not measured" (p. 193).

The literature on oral examinations is less plentiful or detailed, but its general outlines follow the points made about written examinations. Three studies, however, are noteworthy in intimating that with training of examiners, suitable goals, and use of appropriate ratings scales, orals can be made sufficiently reliable. Butzin et al. (1982) reports on the satisfactory reliability of the grading process used in the American Board of Pediatrics oral examination. James Frith (1979) comments favorably on the reliability evidenced in field-tests of the U.S. State

Department's Foreign Service Institute (FSI) criterion-referenced testing system, normally used to evaluate the language competences of Foreign Service and other U.S. government representatives, but effective in the pilot project in assessing the language skills of college students. Finally, Granville Johnson (1972) presents a ratings scale for master's orals that could serve as a point of departure for devising other scales to score undergraduate oral presentations and examinations.

American applications

Most faculty are already accustomed to external examiners for purposes of institutional reaccreditation and departmental self-study. As described by a provost at Ohio State (Adelman, 1985, p. 55):

> The external review team makes judgments on the appropriateness of the program's goals, the quality of the curriculum, and the effectiveness of the faculty effort in research and teaching. The external reviewers identify weaknesses and make recommendations.

The judgments of such reviewers have at least suasive power and, in some cases, can influence reaccreditation and funding decisions. Nevertheless, learning is only indirectly assessed in these processes, whereas in the examples that follow, external examiners are directly involved in assessing student outcomes.

On the course level, Thomas Sawyer (1976) of the University of Michigan reported successful use of externals in a technical writing course taken by senior engineering majors to satisfy a senior rhetoric requirement. In addition to weekly papers, students were asked to prepare a report normally required in an engineering design or research course taken the same term. The same report was submitted to both the content course and the writing course, but in the latter, the report and a ten-minute oral presentation on it were graded by two external examiners, one an engineer and the other an English professor. The professors were drawn from departments of neighboring universities; the engineers from local industries. Each examiner wrote brief comments and recommended a final course grade. Sawyer reported that "in more than half of the cases they have both recommended the identical grades and in most of the other cases they have differed by only one grade" (p. 345). The benefit to students was that they were asked to perform a task, writing and delivering a report on engineering options, that would be a large part of their work after graduation, and they received constructive feedback (and the written comments) from professionals both in the discipline and in writing. Sawyer also stressed that a concomitant benefit to the instructor was that the process served as an external check on teaching--directly in the case of his own course and indirectly on the substantive education the senior engineering majors had received.

A model for independent study is that pioneered by the National Association for Self-Instructional Language Programs (SILP). Designed to offer students opportunity for guided individualized study in foreign languages not regularly offered in the institutional curriculum, SILP involves contractual relationships between students and a campus coordinator who can provide resources such as written and taped materials and, in particular, arrange for locally-available native speakers of the language to serve as tutors. Of immediate interest is the following description from Coffin (1975, p. 2):

> . . . a qualified outside examiner. . .checks and evaluates the students' work at the end of each term of learning. The outside examiner is a person who has the proper academic credentials and qualifications and who is actively engaged in the instruction of the target language in a recognized academic institution. In order to complete the course of study and to receive academic credit for work done during the term, students must be examined at the end of the term by this language specialist. It is not only a process of evaluation of individual students' work, but it is also a process by which the standards and quality of the program are maintained.

The final examination is an oral examination, since oral proficiency is normally the aim of students. Coffin notes that examiners need to be language professionals who are familiar with SILP and who can advise from the outset with regard to choice of materials, particulars of learning contracts, and appropriateness of on-site tutors. Morehouse and Boyd-Bowman (1973, p. 7), in a multi-institutional study of SILP, emphasize that:

> the examiners play a key troubleshooting or technical assistance role. Being experienced teachers of the languages being studied through the self-instructional program at neighboring universities, they are able to diagnose problems with individual students and their tutors which campus coordinators, ordinarily being unfamiliar with the languages being studied, cannot do.

Alverno College has been justly renowned for its college-wide approach to assessment. One particular dimension of its program is its use of local professionals and employers as on-campus examiners and as off-campus assessors for student internships (Mentkowski and Doherty, 1983). A significant requirement is that prospective assessors must attend a series of sessions designed to orient them to the college's purposes and assessment techniques. Alverno's emphasis on feedback to students in order to facilitate performance demands that any assessor must not simply grade, but also must provide both

constructive criticism and extensive debriefing for students to demonstrate learning.

American uses of examiners go beyond individual courses to encompass the more traditional area of advanced work and summative learning in the major. Many colleges which require senior shows in art, music, or dance have invited artists and critics to be members of evaluation committees. Senior theses and other types of honors projects have also been opportunities to use external examiners. Knox College, for one, requires that honors candidates complete research projects and undergo an oral defense before faculty external to the institution. King College has provisions whereby departmental honors candidates meet with external examiners, drawn from other institutions or from local professionals with appropriate backgrounds, for half-hour oral interviews. An ancillary benefit of such encounters has been the offer of jobs or fellowships for graduate study.

The best-known example of an ongoing external examiner system at an American college, however, is that of Swarthmore. Established in 1923, the system is oriented towards the college's honors candidates, presently about one-third of the undergraduates. The designation of honors depends solely upon performance in the senior year on a series of comprehensive examinations prepared and graded by external faculty examiners. Students undertake a plan of study beginning in the junior year and negotiate a series of seminars. The terminal examinations are actually in the seminar fields, not in the entirety of the discipline. Nevertheless, Jones (1933, pp. 96-7) found:

> Whatever the inadequacies of outside examining may be, the possibility of poor examiners, the occasional bias or inadequate sampling of questions, and the possibility that some students might be too emotionally distraught to represent their true abilities--all these have been more than compensated for by the tremendously free spirit of cooperation and inquiry that has been developed. The teacher is a guide and a companion, an honest critic of the student's ability. But he is not to be the final judge. The student and teacher are both to be judged by an outside court.

A 1967 evaluation of the program (quoted in Milton and Edgerly, 1976, p. 53), reiterated:

> Many external examiners. . .think the system works well, and the examiners' evaluations of students are generally consistent with the faculty's. Many graduates of honors have said, as have many faculty, that the system helps to create an

> atmosphere of faculty-student collaboration. . . . These are now conventional statements; but we are inclined to agree with them. The colleagueship and the intellectual checks provided by external examiners are widely felt to be valuable for both students and the faculty; many of the latter, especially, set high store by it.

The continuing vitality of the Swarthmore model was an inspiration behind the Association of American College's pilot program, "Assessing Learning in Academic Majors Using External Faculty Examiners" (1986). Underwritten by the Fund for the Improvement of Postsecondary Education (FIPSE), the project runs from 1986-89. Eighteen institutions have been clustered in groups of three by characteristics of region and institutional size and by similarity of academic program offerings. Each cluster has designated three majors that its institutions will examine in common. Beginning in 1988, fifteeen students graduating in a given major will be examined through written and oral exercises by faculty in that discipline from the other two institutions of the cluster. Results will be provided to students, and visiting examiners will report to the department and to the institutional officers to help them assess how the objectives of the major are being met.

The project aims to field-test the applicability of external examiners in a variety of educational situations, from large universities to small liberal arts colleges, from the East to the South to the far West. One overriding consideration has been to adapt the arrangements to the ongoing programs of a given institution. Therefore, while most institutions will be using a combination of comprehensive written and oral examinations, several schools with thesis requirements will be substituting these projects for the written portion. Some schools will be selecting students for participation from across the entire cohort of prospective graduates that year in the discipline; others have decided to use only honors candidates. Some institutions are interested in having participating students take a capstone seminar in the discipline; others prefer to see if the general progress through the major sufficiently prepares students in the knowledge and skills considered essential in the field.

The AAC and FIPSE hope that the final report will offer a number of models for the use of external examiners, describe the conditions under which certain models flourished or withered, and encourage other colleges and universities to consider this approach as an option in their search for appropriate means of assessing student learning and program quality.

Final observations

The suppleness of the external examiner approach has much to recommend it. Unlike programs for assessment which may demand an initial wholesale commitment of an institution, the use of externals may begin with a single course or department.

In the current scene, external examiners offer a particularly effective way to assess both student learning and curricular coherence in a major. Faculty expertise is no doubt displayed to best advantage in departmental courses, and grades within courses are probably, on the whole, accurate estimations of student achievement. But the shape of total learning in the discipline is too often left up to the serendipity of student electives. That the sum of courses taken in a major by a student will add up to anything coherent and complete is too often a matter of hope rather than of design. The use of externals, particularly combined with a scheme of comprehensive examinations, whether written or oral or both, can not only probe for cumulative learning patterns but also lead to valuable information and recommendations as to where curricular requirements need to be more specific and how course offerings need to be strengthened.

By no means should external examiner ratings supercede the aggregate judgment of instructors over the course of a student's education. But used, for example, in conjunction with the awarding of academic honors, they can serve as a validation of both student learning and institutional programs for purposes of accountability to external agencies. When combined with course grades and judicious use of objective tests, external examinations can serve as one prong of a multi-dimensional approach to summative evaluation, really the only way of getting close to a complete, accurate, and fair picture of student achievement. In this way, we may benefit from the British experience with externals while avoiding some of that system's drawbacks.

And what of costs? In the American examples I've cited, remuneration to examiners has ranged from coffee and cookies to reimbursement for expenses, from a year-end banquet to honoraria of $50-$100 per term. Some clusters in the AAC/FIPSE project have chosen to institute a capstone seminar in the major as a way of giving a teaching course credit to those who will be serving as external examiners. Instructors will teach their own students in the seminar but prepare and administer comprehensives to seminar students at the other cluster institutions. This exchange of faculty services between schools as part of official teaching loads constitutes institutional recognition of assessment activities as an integral part of instruction and service.

It is worthwhile, however, to note why industry professionals and distinguished academics are presently serving as examiners at rates far below their usual consulting fees. Some cite the opportunity to learn about campus programs and students in order to enhance their own companies and departments. Others see the role as an extension of service to the profession and to the community. At heart, however, is the notion of collegiality, the desire to engage in common cause with peers across institutional boundaries to facilitate the instruction and learning of the next generation of professionals and academics. The human contacts encouraged by the external examiner approach are by no means the least of its potential benefits.

REFERENCES

Adelman, Clifford, ed. From Report to Response: Proceedings of Regional Conferences on the Quality of American Higher Education. Washington: U.S. Department of Education, 1985.

Agar, Margaret and Judith Weltman. "The Present Structure of University Examinations." Universities Quarterly, 21:3 (June 1967), 272-85.

Association of American Colleges. Assessing Learning in Academic Majors Using External Faculty Examiners. Grant proposal to the Fund for the Improvement of Postsecondary Education. Washington: Association of American Colleges, 1986.

Astin, Alexander. Achieving Educational Excellence. San Francisco: Jossey-Bass, 1985.

Bassey, Michael. The Assessment of Students by Formal Assignments. Wellington: New Zealand Univ. Students Association, 1971. (ED 059654)

Bloom, Benjamin S., et al. Taxonomy of Educational Objectives. London: Longmans, 1956.

Butzin, Diane W., et al. "A Study of the Reliability of the Grading Process Used in the American Board of Pediatrics Oral Examination." Journal of Medical Education, 57:12 (December 1982), 944-46.

Cockburn, Barbara and Alec Ross. Inside Assessment. Teaching in Higher Education Series, 7. Lancaster, Eng.: Lancaster Univ., 1980. (ED 230104)

Coffin, Edna Amir. "Preliminaries and Preparation for Examinations: Examiner's Report and Evaluation." Paper presented at the Conference on the National Association for Self-Instructional Language Programs, 19-20 September 1975. (ED 119505)

Committee of Vice Chancellors and Principals. "External Examiner System for First Degree and Taught Master's Courses: Code of Practice." Pamphlet, published Univ. of Liverpool, 1984.

Cox, Roy. "Examinations and Higher Education: A Survey of the Literature." Universities Quarterly, 21:3 (June 1967), 292-340.

Cronbach, Lee J. "Test Validation." In Educational Measurement, ed. Robert L. Thorndike. 2nd edition. Washington: American Council on Education, 1971, pp. 443-507.

Dale, R. R. "University Standards." Universities Quarterly, 13:2 (February 1959), 186-95.

Eells, Walter Crosby. "Reliability of Repeated Grading of Essay Type Examinations." Journal of Educational Psychology, 21 (1930), 48-52.

Elton, Lewis. "Assessment for Learning." In Professionalism and Flexibility in Learning, ed. Donald Bligh. Surrey, Eng.: Society for Research in Higher Education, 1982, pp. 106-35.

Frederiksen, Norman. "The Real Test Bias: Influences of Testing on Teaching and Learning." American Psychologist, 39:3 (March 1984), 193-202.

Frith, James R. "Testing the FSI Testing Kit." ADFL Bulletin, 11:2 (November 1979), 12-4.

Harris, John. "Assessing Outcomes in Higher Education." In Assessment in American Higher Education, ed. Clifford Adelman. Washington: U.S. Department of Education, 1986, pp. 13-31.

Hartog, Phillip and E. C. Rhodes. An Examination of Examinations. London: Macmillan, 1935.

Heywood, John. Assessment in Higher Education. London: John Wiley, 1977.

Johnson, Granville B. "A Rating Scale for Master's Orals." Unpublished paper, 1972. (ED 068582)

Jones, Edward Safford. Comprehensive Examinations in American Colleges. New York: Macmillan, 1933.

Lowell, A. Lawrence. "Examination by Subjects Instead of by Courses." The Harvard Graduates' Magazine, 20:80 (June 1912), 583-92.

Mentkowski, Marcia and Austin Doherty. Careering After College: Establishing the Validity of Abilities Learned in College for Later Careering and Professional Performance. Final report to the National Institute for Education. Milwaukee: Alverno Productions, 1983. (ED 239556)

Milton, Ohmer and John W. Edgerly. The Testing and Grading of Students. New Rochelle, N. Y.: Change Magazine, 1976. (ED 125504)

Morehouse, Ward and Peter Boyd-Bowman. Independent Study of Critical Languages in Undergraduate Colleges. Washington: Institute of International Studies, 1973. (ED 107096)

Omolewa, Michael. "The Rationale for the Use of External Examiners in the Conduct of University Examinations: The Case of the University of Ibadan." In Overseas Universities, ed. R. G. Harris. London: Inter-University Council for Higher Education Overseas, 1978, pp. 19-24. (ED 177934)

O'Neill, Joseph P. "Examinations and Quality Control." In Meeting the New Demands for Standards, ed. Jonathan Warren. New Directions for Higher Education series, no. 43. San Francisco: Jossey-Bass, 1983, pp. 69-79.

Oppenheim, A. N., Marie Jahoda, and R. L. James. "Assumptions Underlying the Use of University Examinations." Universities Quarterly, 21:3 (June 1967), 341-51.

Piper, David Warren. "Enquiry into the Role of External Examiners." Studies in Higher Education, 10:3 (1985), 331-42.

Rippere, Victoria L. "On the 'Validity' of University Examinations: Some Comments on the Language of the Debate." Universities Quarterly, 28:2 (Spring 1974), 209-18.

Roach, J. Public Examinations in England, 1850-1900. London: Cambridge U. P., 1971.

Rowntree, Derek. Assessing Students: How Shall We Know Them? London: Harper and Row, 1977.

Rudolph, Frederick. Curriculum: A History of the American Undergraduate Course of Study Since 1636. San Francisco: Jossey-Bass, 1977.

Sawyer, Thomas M. "External Examiners: Separating Teaching From Grading." Engineering Education, 66:4 (January 1976), 344-6.

Warren, Jonathan R. "Standardized Open-Ended Questions as Measures of General Educational Goals." Paper presented at the Third International Symposium on Educational Testing, Leyden, The Netherlands, 30 June 1977.

Williams, W. F. "The Role of the External Examiner in First Degrees." Studies in Higher Education, 4:2 (October 1979), 161-8.

ADDITIONAL RESOURCES

AAHE ASSESSMENT FORUM

The following resources are available for purchase from the AAHE Assessment Forum:

1. Resource Packet: Five Papers **$15.00**

--"Assessment, Accountability, and Improvement: Managing the Contradiction," P. Ewell
--"Assessment and Outcomes Measurement: A View from the States," C. Boyer, P. Ewell, J. Finney, J. Mingle
--"The External Examiner Approach to Assessment," B. Fong
--"Six Stories: Implementing Successful Assessment," P. Hutchings
--"Thinking About Assessment: Perspectives for Presidents and Chief Academic Officers," E. El-Khawas and J. Rossmann

2. Three Presentations: **$8.00**

from the 2nd National Conference on Assessment in Higher Education
--Lee S. Shulman -- "Assessing Content and Process: Challenges for the New Assessments"
--Virginia B. Smith -- "In the Eye of the Beholder: Perspectives on Quality"
--Donald M. Stewart -- "The Ethics of Assessment"

3. Audio Tape: Shulman Address **$9.00**

"Assessing Content and Process: Challenges for the New Assessments," Lee S. Shulman

4. Address Roster of Denver Conference Participants **$3.00**

Available Soon:

National Directory on Assessment in Higher Education

To order items indicated above, and for more information about future Assessment Forum resources, services, and activities, contact: Patricia Hutchings, Director, AAHE Assessment Forum, One Dupont Circle, Suite 600, Washington, DC 20036; 202/293-6440.

Orders under $25 must be prepaid. Allow four weeks for delivery. Postage and handling is included in prices quoted.

MAKE CHECKS PAYABLE TO AAHE ASSESSMENT FORUM